Jurnal Minfo Polgan (JMP)
Volume 15, Nomor 3, September-Desember 2026
DOI : https://doi.org/10.33395/jmp.v15i3.16698
e-ISSN : 2797-3298
p-ISSN : 2089-9424

Perbandingan Kinerja Algoritma Random Forest, XGBoost, dan CatBoost untuk Klasifikasi Risiko Diabetes

[1]Nurhajar Anugraha, [2]Suzan Zefi, [3]Emilia Hesti, [4]Sudirman Yahya, [5]Muhammad Surya Ragasin, [6]Muhammad Daffa, [7]M. Fathur Rahman Al-Faqiih
[1*,2,3,4,5,6,7]Pendidikan Sarjana Terapan pada JurusanTeknik Elektro Program Studi Teknik Telekomunikasi, Politeknik Negeri Sriwijaya, Palembang, Indonesia
*Korespondensi: nurhajar.anugraha@polsri.ac.id

Submit : 30 Jul 2026 | **Diterima** : 27 Agust 2026 | **Terbit** : 10 Sept 2026

ABSTRACT

This study focuses on a comparative performance analysis of Random Forest, XGBoost, and CatBoost in predicting diabetes risk using the "Health and Lifestyle Data for Diabetes Prediction" dataset from Kaggle. A quantitative research approach was employed, comprising several stages: data collection and preprocessing, selection of target variables and features, and model training and testing. The dataset was partitioned using the Stratified Train-Test Split method with an 80:20 ratio to maintain class distribution across the training and testing sets. Subsequently, the performance of each algorithm was evaluated based on accuracy, precision, recall, F1-score, and the confusion matrix. The results indicate that all the algorithms achieved classification accuracy levels exceeding 91%. XGBoost demonstrated the best performance with an accuracy of 91.61%, followed by Random Forest at 91.45% and CatBoost at 91.34%. Metrics for precision, recall, and F1-score also revealed relatively balanced classification performance across the three algorithms. These findings demonstrate that ensemble learning methods are effective at identifying diabetes risk by leveraging health and lifestyle characteristics. Among the algorithms compared, XGBoost yielded the most optimal results for the dataset used, making it the most suitable model for this study. This research serves as a reference for developing diabetes risk prediction systems using machine learning approaches. Future work could involve hyperparameter optimization, cross-validation techniques, and the use of datasets with more diverse characteristics to enhance the models' generalization capabilities.

Keywords: *CatBoost; Diabetes; Ensemble Learning; Machine Learning; Risk Prediction; Random Forest; XGBoost.*

ABSTRAK

Penelitian ini berfokus pada analisis perbandingan performa Random Forest, XGBoost, dan CatBoost dalam melakukan prediksi risiko diabetes dengan menggunakan dataset Health and Lifestyle Data for Diabetes Prediction dari Kaggle. Metode penelitian yang digunakan adalah pendekatan kuantitatif yang terdiri atas beberapa proses, mulai dari pengumpulan dan preprocessing data, pemilihan variabel target serta fitur, hingga proses pelatihan dan pengujian model. Dataset dibagi dengan metode Stratified Train-Test Split menggunakan proporsi 80:20 untuk menjaga distribusi kelas pada data latih dan data uji. Selanjutnya, performa masing-masing algoritma dianalisis berdasarkan accuracy, precision, recall, F1-score, serta confusion matrix. Hasil pengujian memperlihatkan bahwa seluruh algoritma yang digunakan mampu memberikan hasil klasifikasi dengan tingkat akurasi di atas 91%. XGBoost menghasilkan performa terbaik dengan akurasi 91,61%, sedangkan Random Forest mencapai 91,45% dan CatBoost sebesar 91,34%. Hasil pengukuran precision, recall, dan F1-score juga memperlihatkan performa klasifikasi yang relatif seimbang pada ketiga algoritma. Berdasarkan hasil tersebut, metode ensemble learning memiliki kemampuan yang baik dalam mengidentifikasi risiko diabetes dengan memanfaatkan karakteristik kesehatan dan gaya hidup. Di antara algoritma yang dibandingkan, XGBoost memberikan hasil paling optimal pada dataset yang digunakan sehingga dapat dipertimbangkan sebagai model yang paling sesuai dalam penelitian ini. Penelitian ini dapat menjadi bahan acuan untuk pengembangan sistem prediksi risiko diabetes dengan pendekatan Machine Learning. Pengembangan lebih lanjut dapat dilakukan dengan menerapkan optimasi hyperparameter, teknik cross-validation, dan dataset

dengan karakteristik yang lebih beragam agar model memiliki kemampuan generalisasi yang lebih baik.

Kata Kunci: CatBoost; Diabetes; Ensemble Learning; Machine Learning; Prediksi Risiko; Random Forest; XGBoost.

PENDAHULUAN

Diabetes melitus menjadi salah satu permasalahan kesehatan yang perlu mendapat perhatian karena jumlah penderitanya terus bertambah secara global. Kondisi tersebut berkaitan dengan keadaan ketika kadar glukosa dalam darah berada di atas batas normal sebagai akibat adanya gangguan pada produksi insulin maupun respons tubuh terhadap insulin. Apabila tidak memperoleh penanganan yang sesuai, diabetes dapat berkembang menjadi berbagai komplikasi, antara lain gangguan pada sistem kardiovaskular, fungsi ginjal, saraf, dan retina, serta dapat meningkatkan risiko kematian (Saleh et al., 2024). Kondisi tersebut menunjukkan pentingnya identifikasi risiko diabetes sejak tahap awal sehingga tindakan pencegahan dan pengendalian dapat dilakukan sebelum kondisi berkembang menjadi lebih serius. Deteksi risiko secara dini juga memungkinkan pemberian penanganan yang lebih cepat sesuai dengan kondisi individu (Afolabi et al., 2025)

Kemajuan teknologi *Artificial Intelligence* (AI) telah mendorong pemanfaatannya dalam berbagai permasalahan, termasuk dalam pengolahan dan analisis data kesehatan. Salah satu teknologi yang berada dalam lingkup AI adalah *Machine Learning*, yang bekerja dengan memanfaatkan pola yang terdapat pada data untuk membangun kemampuan prediksi tanpa harus menentukan seluruh aturan secara manual melalui pemrograma (Azizah et al., 2025), (Md Ashraful Alam ,Amir Sohel, Kh Maksudul Hasan, 2024). Penerapan teknologi tersebut dalam bidang kesehatan telah mencakup berbagai kebutuhan, seperti membantu identifikasi penyakit, memperkirakan risiko suatu kondisi kesehatan, hingga mendukung proses pengambilan keputusan. Pemanfaatan data pasien melalui *Machine Learning* memungkinkan proses analisis dilakukan secara lebih efisien serta dapat membantu meningkatkan ketepatan hasil prediksi.

Pemanfaatan *Machine Learning* untuk permasalahan diabetes telah dikaji melalui berbagai pendekatan dan algoritma (Andini et al., 2026; Ibrahim et al., 2025). (Tripathy et al., n.d.) membandingkan beberapa algoritma *Machine Learning* dan *Deep Learning* pada prediksi diabetes serta menunjukkan bahwa pendekatan *Deep Learning* menghasilkan kinerja yang lebih unggul, meskipun algoritma *Machine Learning* masih memberikan hasil yang kompetitif pada dataset kesehatan. Penelitian (Afolabi et al., 2025) membandingkan algoritma *Logistic Regression, Random Forest,* dan K-*Nearest Neighbor* menggunakan data *electronic health records*, di mana K-*Nearest Neighbor* memberikan kinerja terbaik pada dataset yang digunakan. Sementara itu, penelitian (Jithendra et al., 2012) membandingkan beberapa algoritma klasifikasi dan menunjukkan bahwa pendekatan *Machine Learning* mampu menghasilkan model prediksi yang efektif sebagai alat bantu deteksi dini diabetes.

Seiring berkembangnya penelitian, pendekatan *ensemble learning* semakin banyak diterapkan karena mampu meningkatkan stabilitas dan akurasi model dibandingkan algoritma tunggal. (Kharis Syaban, 2025) menunjukkan bahwa berbagai metode *ensemble learning* mampu meningkatkan kinerja prediksi risiko diabetes dibandingkan model dasar. (Dip Das, Aayushman, Sourav Kumar, Md Amir Hussain, 2025) juga membandingkan beberapa algoritma *ensemble learning*, yaitu XGBoost, *Random Forest*, CatBoost, LightGBM, dan KNN pada dataset BRFSS, serta memperoleh hasil prediksi yang sangat baik setelah dilakukan penanganan ketidakseimbangan data menggunakan SMOTE-ENN. Selain itu, (Irfannandhy et al., 2024) membandingkan *Random Forest* dan CatBoost dengan teknik SMOTE. Hasilnya menunjukkan bahwa kedua algoritma tersebut dapat digunakan untuk memprediksi risiko diabetes dengan performa yang baik, walaupun nilai *precision, recall, F1-score*, dan akurasi yang diperoleh tidak sepenuhnya sama.

Selain pemilihan algoritma, tahapan *preprocessing* juga menjadi faktor penting dalam membangun model prediksi yang optimal. (Linkon et al., 2024) menunjukkan bahwa transformasi fitur dan teknik *preprocessing* yang tepat dapat meningkatkan performa berbagai model *Machine Learning* pada deteksi dini diabetes. Setelah proses pelatihan model dilakukan, evaluasi menggunakan metrik seperti *accuracy*, *precision*, *recall*, *F1-score*, serta *confusion matrix* diperlukan untuk memberikan gambaran yang lebih lengkap mengenai kemampuan model dalam membedakan kelas pada proses klasifikasi (Alshammari, 2024).

Meskipun berbagai penelitian telah mengembangkan model prediksi diabetes

menggunakan algoritma Machine Learning maupun ensemble learning, masih terdapat beberapa keterbatasan dalam aspek perbandingan model. Penelitian sebelumnya telah menunjukkan bahwa Random Forest dan CatBoost mampu memberikan performa yang baik dalam prediksi risiko diabetes, terutama dengan penerapan teknik penyeimbangan data seperti SMOTE. Namun, penelitian tersebut belum membandingkan Random Forest, XGBoost, dan CatBoost secara bersamaan pada dataset kesehatan dan gaya hidup yang sama. Selain itu, penelitian yang menggunakan pendekatan ensemble masih memiliki perbedaan dalam pemilihan dataset, tahapan preprocessing, dan metode evaluasi, sehingga hasil kinerja antar algoritma belum dapat dibandingkan secara langsung.

Berdasarkan celah penelitian tersebut, penelitian ini dilakukan untuk mengevaluasi dan membandingkan kemampuan *Random Forest*, XGBoost, dan CatBoost dalam memprediksi risiko diabetes menggunakan dataset *Health and Lifestyle Data for Diabetes Prediction* yang tersedia pada platform Kaggle (Shihab, 2026). Berbeda dari penelitian yang hanya berfokus pada satu atau dua algoritma, penelitian ini menempatkan ketiga metode *ensemble learning* tersebut dalam skenario pengujian yang sama. Target klasifikasi yang digunakan mencakup tiga kondisi, yaitu *No Diabetes, Pre-Diabetes*, dan *Type 2 Diabetes*. Kinerja setiap model kemudian dianalisis melalui *accuracy, precision, recall, F1-score*, serta *confusion matrix*. Dengan demikian, hasil penelitian tidak hanya digunakan untuk melihat model dengan tingkat akurasi tertinggi, tetapi juga untuk memperoleh gambaran perbandingan performa masing-masing algoritma secara lebih menyeluruh dan menentukan pendekatan yang paling sesuai dengan karakteristik dataset yang digunakan.

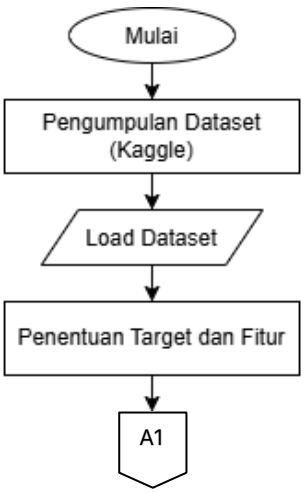
METODE PENELITIAN

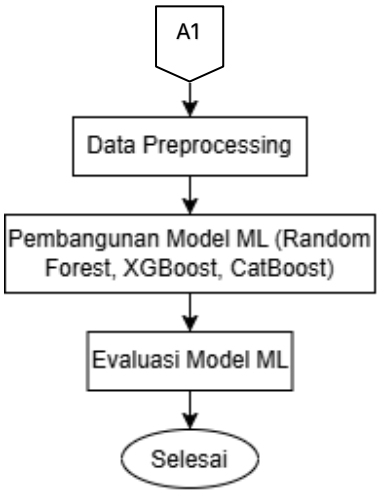
Penelitian ini menggunakan pendekatan kuantitatif dengan metode *Machine Learning* untuk membandingkan Random Forest, XGBoost, dan CatBoost dalam prediksi risiko diabetes. Dataset *Health and Lifestyle Data for Diabetes Prediction* diperoleh dari Kaggle dan terdiri atas 97.297 data dengan 31 variabel. Target yang digunakan adalah diabetes_stage dengan tiga kelas, yaitu *No Diabetes, Pre-Diabetes*, dan *Type 2 Diabetes*. Fitur penelitian meliputi variabel kesehatan dan gaya hidup seperti usia, jenis kelamin, BMI, status merokok, aktivitas fisik, riwayat diabetes keluarga, dan *diet score*.

Sebelum pemodelan, dilakukan pemeriksaan tipe data, duplikasi, dan *missing value*. Tidak ditemukan *missing value* pada fitur yang digunakan. Fitur kategorikal diproses menggunakan *One-Hot Encoding*, sedangkan fitur numerik menggunakan *StandardScaler*. Seluruh preprocessing diterapkan melalui *Pipeline* dan *ColumnTransformer* untuk menjaga konsistensi pengolahan data serta mengurangi *data leakage*.

Dataset dibagi dengan rasio 80:20 menggunakan *stratified train-test split* berdasarkan *diabetes_stage* dan random_state=42. Model Random Forest, XGBoost, dan CatBoost kemudian dilatih menggunakan parameter yang telah ditentukan tanpa *hyperparameter tuning*. Ketidakseimbangan kelas ditangani dengan *class weighting* pada Random Forest dan CatBoost. Evaluasi dilakukan menggunakan *accuracy, precision, recall, F1-score*, dan *confusion matrix* untuk menentukan model dengan performa tertinggi.

Tahapan penelitian yang dilakukan mulai dari pengumpulan dataset hingga evaluasi performa model disajikan pada Gambar 1.





Gambar 1. Tahapan Metode Penelitian

1. Mulai
   Penelitian diawali dengan proses identifikasi permasalahan mengenai pentingnya deteksi dini risiko diabetes serta penentuan metode penelitian yang digunakan. Selanjutnya disusun tahapan penelitian mulai dari pengumpulan dataset hingga evaluasi performa model machine learning.
2. Pengumpulan Dataset (Kaggle)
   Tahap awal penelitian dilakukan dengan mengumpulkan dataset Health and Lifestyle Data for Diabetes Prediction yang diperoleh dari platform Kaggle (Shihab, 2026). Dataset ini dipilih karena memuat berbagai informasi yang berkaitan dengan karakteristik individu, gaya hidup, dan riwayat kesehatan yang berpengaruh terhadap risiko diabetes. Penggunaan dataset tersebut memungkinkan pembangunan model prediksi tanpa memerlukan data hasil pemeriksaan laboratorium sehingga lebih praktis untuk penelitian prediksi berbasis *machine learning.*
3. Load Dataset
   Dataset yang telah diperoleh kemudian dimuat ke dalam lingkungan pemrograman Python menggunakan pustaka Pandas. Pada tahap ini dilakukan pembacaan dataset ke dalam bentuk *DataFrame* sehingga data dapat diakses, divisualisasikan, dan diolah pada tahapan selanjutnya.
4. Penentuan Target dan Fitur
   Tahap ini bertujuan menentukan variabel target dan fitur yang digunakan dalam proses pemodelan. Variabel target penelitian adalah diabetes_stage, yang terdiri atas tiga kelas, yaitu No Diabetes, Pre-Diabetes, dan Type 2 Diabetes. Sementara itu, fitur yang digunakan meliputi variabel demografi, gaya hidup, serta riwayat kesehatan yang tersedia pada dataset, seperti usia, jenis kelamin, *Body Mass Index* (BMI), aktivitas fisik, pola makan, kebiasaan merokok, konsumsi alkohol, riwayat keluarga diabetes, serta atribut lain yang relevan terhadap prediksi risiko diabetes.
5. *Data Preprocessing*
   Sebelum digunakan dalam pemodelan, dataset terlebih dahulu melalui tahap *data preprocessing* untuk memastikan data berada dalam kondisi yang sesuai. Proses ini diawali dengan pengecekan kondisi dan kualitas data, termasuk identifikasi *missing value* serta penanganannya jika terdapat data yang tidak lengkap. Variabel yang bersifat kategorikal kemudian dikonversi ke dalam bentuk numerik melalui metode *Label Encoding*. Selanjutnya, fitur numerik dilakukan standardisasi menggunakan *StandardScaler*. Setelah seluruh tahapan transformasi selesai, dataset dibagi menjadi data latih dan data uji dengan rasio 80:20 menggunakan metode *Stratified Train-Test Split*. Pembagian secara stratifikasi dilakukan agar proporsi kelas pada data latih dan data uji tetap terjaga. Seluruh tahapan preprocessing diterapkan dengan prosedur yang sama sehingga data yang digunakan oleh ketiga algoritma memiliki karakteristik pengolahan yang konsisten.

6. Pembangunan Model *Machine Learning*
   Setelah data siap digunakan, tahap berikutnya adalah membangun model klasifikasi dengan menerapkan tiga algoritma *ensemble learning*, yaitu Random Forest, XGBoost, dan CatBoost. Pemilihan ketiga algoritma tersebut didasarkan pada kemampuannya dalam mengolah data berbentuk tabular serta menangkap pola hubungan yang bersifat nonlinier antarfitur. Selain itu, algoritma tersebut banyak digunakan dalam permasalahan klasifikasi karena mampu memberikan hasil prediksi yang baik. Pada tahap ini, masing-masing algoritma dilatih menggunakan data latih yang telah diproses sebelumnya untuk membentuk model yang dapat digunakan dalam memprediksi tingkat risiko diabetes.
7. Evaluasi Model *Machine Learning*
   Model yang telah dibangun selanjutnya diuji untuk mengetahui tingkat keberhasilannya dalam melakukan klasifikasi. Pengujian dilakukan menggunakan data uji yang tidak digunakan pada proses pelatihan. Performa setiap model dianalisis berdasarkan beberapa metrik evaluasi, meliputi *accuracy, precision, recall*, F1-*score*, serta *confusion matrix*. Nilai dari masing-masing metrik kemudian digunakan sebagai dasar untuk membandingkan kemampuan Random Forest, XGBoost, dan CatBoost dalam mengelompokkan data berdasarkan risiko diabetes. Melalui hasil evaluasi tersebut dapat ditentukan algoritma yang memberikan kinerja paling optimal ketika diterapkan pada dataset *Health and Lifestyle Data for Diabetes Prediction.*
8. Selesai
   Penelitian diakhiri setelah seluruh model berhasil dievaluasi dan dilakukan analisis terhadap hasil perbandingan performa ketiga algoritma.

HASIL DAN PEMBAHASAN

Pembahasan

Hasil penelitian diperoleh dari proses pembangunan dan evaluasi model *machine learning* untuk prediksi risiko diabetes. Tahapan yang dibahas meliputi pengumpulan dataset, *preprocessing* data, pembangunan model menggunakan algoritma Random Forest, XGBoost, dan CatBoost, serta evaluasi performa masing-masing model berdasarkan metrik *accuracy*, *precision*, *recall*, *F1-score*, dan *confusion matrix*. Hasil evaluasi tersebut kemudian dianalisis untuk membandingkan kinerja ketiga algoritma dalam memprediksi risiko diabetes.

Pengumpulan Dataset

Data penelitian bersumber dari dataset "Health and Lifestyle Data for Diabetes Prediction" yang tersedia pada platform Kaggle. Dataset ini berisi lebih dari 1.000 data individu dengan lebih dari 30 fitur yang mencakup data demografis, gaya hidup, riwayat kesehatan, serta parameter klinis (Shihab, 2026). Label yang tersedia dalam dataset meliputi:

1. diagnosed_diabetes (klasifikasi biner), merupakan label dengan tipe klasifikasi biner, yang menunjukkan apakah seseorang telah didiagnosis menderita diabetes atau tidak. Nilai label ini umumnya direpresentasikan dalam bentuk 0 (tidak diabetes) dan 1 (diabetes),
2. diabetes_stage (klasifikasi multikelas), merupakan label dengan tipe klasifikasi multikelas yang memberikan informasi lebih detail mengenai kondisi diabetes seseorang. Label ini tidak hanya menunjukkan keberadaan diabetes, tetapi juga membedakan tahapan kondisi kesehatan, seperti No Diabetes, Pre-Diabetes, dan Type 2 Diabetes. Dengan adanya pembagian tahapan ini, model prediksi dapat memberikan hasil yang lebih informatif dan mendukung pendekatan pencegahan serta pemantauan risiko secara bertahap,
3. diabetes_risk_score (regresi), merupakan label dengan tipe regresi, yang merepresentasikan tingkat risiko diabetes dalam bentuk nilai numerik kontinu. Label ini umumnya digunakan untuk analisis kuantitatif risiko, bukan untuk klasifikasi kategori kesehatan secara langsung.

Data Preprocessing

Pada tahap ini, dilakukan serangkaian proses persiapan dataset untuk memastikan data layak dan siap digunakan dalam pembangunan model *Machine Learning.*

1. Tahap pertama adalah proses *import pustaka (library)* yang diperlukan untuk mendukung seluruh tahapan pengolahan data, pembangunan model, evaluasi, serta visualisasi hasil.

```python
# Import Pustaka
import os
import joblib
import pickle
import pandas as pd
import numpy as np

# Pustaka Visualisasi
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.metrics import ConfusionMatrixDisplay, confusion_matrix

# Pustaka ML
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder, StandardScaler, OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from catboost import CatBoostClassifier
from xgboost import XGBClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, classification_report
```

Gambar 2. Import Pustaka

Pustaka seperti pandas dan NumPy digunakan pengolahan data, sedangkan matplotlib dan seaborn dimanfaatkan untuk visualisasi performa model, termasuk confusion matrix. Selain itu, pustaka dari scikit-learn digunakan untuk proses preprocessing data, pembagian data latih dan data uji, serta evaluasi kinerja model. Beberapa algoritma klasifikasi seperti CatBoost, XGBoost, dan Random Forest diimpor untuk keperluan perbandingan performa. Proses import pustaka ini bertujuan untuk memastikan seluruh kebutuhan komputasi dan analisis tersedia sebelum proses pelatihan dan pembentukan model Machine Learning.

2. Load Dataset

Pada tahap ini, dataset dimuat ke dalam sistem menggunakan pustaka pandas dari file berformat CSV dan disimpan dalam bentuk DataFrame agar mudah diolah.

```
# Bagian 2: Pemuatan Data, Pemisahan X & y, dan Analisis Distribusi Target

df = pd.read_csv("/content/drive/MyDrive/Final Project Diabetes Flask/Diabetes_and_LifeStyle_Dataset .csv")
df.head()
```

	Age	gender	ethnicity	education_level	income_level	employment_status	smoking_status	alcohol_consumption_per_week	physical_activity_minutes_per
0	58	Male	Asian	Highschool	Lower-Middle	Employed	Never	0	
1	52	Female	White	Highschool	Middle	Employed	Former	1	
2	60	Male	Hispanic	Highschool	Middle	Unemployed	Never	1	
3	74	Female	Black	Highschool	Low	Retired	Never	0	
4	46	Male	White	Graduate	Middle	Retired	Never	1	

5 rows × 31 columns

Gambar 3. Load Dataset

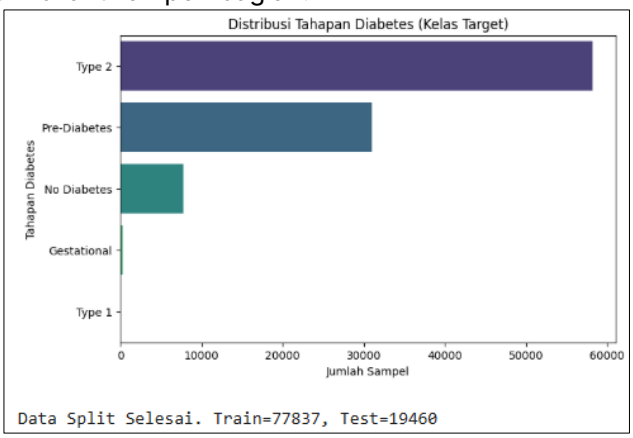
3. Penentuan Variabel Target dan Seleksi Fitur

Variabel target yang digunakan adalah kelas diabetes_stage, yang merepresentasikan tahapan kondisi diabetes secara multikelas. Pemilihan target ini bertujuan agar model tidak hanya mengidentifikasi ada atau tidaknya diabetes, tetapi juga memberikan gambaran tingkat risikonya secara lebih informatif.

```python
# ----------------------------------------------------
# 1. DEFINISI FITUR DAN TARGET
# ----------------------------------------------------
TARGET_COLUMN = "diabetes_stage"

SELECTED_FEATURES = [
    "Age",
    "gender",
    "bmi",
    "smoking_status",
    "alcohol_consumption_per_week",
    "physical_activity_minutes_per_week",
    "family_history_diabetes",
    "employment_status",
    "diet_score"
]

X = df[SELECTED_FEATURES]
y = df[TARGET_COLUMN]
```

Gambar 4. Penentuan Variabel Target dan Fitur

4. Visualisasi Distribusi Target
   Tahap ini dilakukan untuk melihat penyebaran data pada masing-masing kelas yang terdapat dalam variabel target. Visualisasi digunakan untuk mengetahui proporsi setiap kategori pada tahapan diabetes sebelum data digunakan dalam proses pemodelan. Variabel target selanjutnya diubah menjadi representasi numerik melalui metode Label Encoding sehingga dapat diterima oleh algoritma Machine Learning. Setelah proses tersebut selesai, dataset dipisahkan menjadi data latih dan data uji dengan komposisi masing-masing 80% dan 20%. Proses pembagian menggunakan teknik stratified split dengan variabel target sebagai dasar stratifikasi. Dengan demikian, komposisi setiap kelas pada data latih (y_train) dan data uji (y_test) tetap mempertahankan proporsi yang terdapat pada dataset sebelum dilakukan pembagian.



Gambar 5. Visualisasi Kelas Target

Hasil visualisasi menunjukkan adanya ketimpangan jumlah sampel pada setiap kelas atau kondisi *imbalanced dataset*. Perbedaan jumlah data antar kategori terlihat cukup besar, sehingga kondisi tersebut perlu diperhatikan karena dapat memberikan pengaruh terhadap hasil klasifikasi. Model berpotensi lebih banyak mempelajari kelas yang memiliki jumlah sampel lebih besar dan kurang mengenali kelas dengan jumlah data yang terbatas. Untuk mempertahankan keterwakilan setiap kelas, pembagian dataset dilakukan menggunakan teknik *stratified split* berdasarkan variabel target. Teknik ini menjaga agar persentase masing-masing kelas pada data latih dan data uji tetap mendekati kondisi distribusi dataset secara keseluruhan. Dengan mempertahankan komposisi tersebut, kemungkinan model mengalami kecenderungan terhadap kelas mayoritas dapat diminimalkan. Selain itu, seluruh kategori target tetap memiliki representasi pada data latih maupun data uji sehingga proses pembelajaran dan pengujian model dapat dilakukan secara lebih proporsional.

Pembangunan Pipeline dan Pelatihan Model

Pada tahap ini dilakukan pembangunan pipeline preprocessing dan pelatihan model klasifikasi.

```
# ------------------------------------------------
# 1. PEMBUATAN PREPROCESSING PIPELINE (ColumnTransformer)
# ------------------------------------------------
numerical_transformer = Pipeline(steps=[
    ('scaler', StandardScaler())
])

categorical_transformer = Pipeline(steps=[
    ('onehot', OneHotEncoder(handle_unknown='ignore'))
])

preprocessor = ColumnTransformer(
    transformers=[
        ('num', numerical_transformer, numerical_features),
        ('cat', categorical_transformer, categorical_features)
    ],
    remainder='passthrough'
)
print("Preprocessing Pipeline Selesai Dibuat.")
```

Gambar 6. Pembangunan Preprocessing Pipeline

Proses preprocessing dirancang menggunakan *ColumnTransformer* untuk memisahkan pengolahan antara variabel numerik dan variabel kategorikal. Pada fitur numerik diterapkan proses standardisasi menggunakan *StandardScaler*, sedangkan fitur kategorikal diubah menjadi representasi numerik melalui metode *One-Hot Encoding*. Pemisahan tersebut dilakukan agar masing-masing jenis fitur dapat diproses menggunakan teknik yang sesuai dengan karakteristik datanya. Dengan demikian, data keluaran preprocessing telah berada dalam bentuk yang dapat digunakan oleh algoritma klasifikasi.

Setelah *pipeline preprocessing* terbentuk, selanjutnya dibuat tiga *pipeline* model untuk proses klasifikasi, yaitu menggunakan CatBoost, Random Forest, dan XGBoost. Ketiga model tersebut kemudian digunakan dalam tahap pelatihan dengan data yang telah melalui proses transformasi sebelumnya.

```
# Model Utama (CatBoost)
cat_model = Pipeline(steps=[
    ('preprocessor', preprocessor),
    ('classifier', CatBoostClassifier(
        iterations=500,
        learning_rate=0.1,
        depth=6,
        verbose=0,
        random_state=42,
        auto_class_weights='Balanced'
    ))
])
```

```
# Model Perbandingan – Random Forest
rf_model = Pipeline(steps=[
    ('preprocessor', preprocessor),
    ('classifier', RandomForestClassifier(
        n_estimators=300,
        random_state=42,
        class_weight='balanced'
    ))
])
```

```
# Model Perbandingan – XGBoost
xgb_model = Pipeline(steps=[
    ('preprocessor', preprocessor),
    ('classifier', XGBClassifier(
        n_estimators=300,
        learning_rate=0.1,
        max_depth=6,
        objective='multi:softmax',
        num_class=len(target_encoder.classes_),
        eval_metric='mlogloss',
        random_state=42
    ))
])
```

Gambar 7. Pembangunan tiga model

Pemilihan Random Forest, XGBoost, dan CatBoost dilakukan dengan mempertimbangkan karakteristik dataset yang memiliki pola hubungan kompleks, bersifat nonlinier, serta menunjukkan distribusi kelas yang tidak seimbang. Ketiga metode tersebut termasuk dalam kelompok *ensemble learning* yang menggabungkan beberapa model dasar untuk meningkatkan kemampuan prediksi. Random Forest menggunakan pendekatan *tree ensemble*, sedangkan XGBoost dan CatBoost menerapkan teknik *boosting* dengan membangun model secara bertahap untuk memperbaiki kesalahan pada tahap sebelumnya. Pendekatan tersebut dinilai sesuai untuk permasalahan prediksi diabetes karena dapat mengenali pola hubungan antarvariabel yang kompleks serta menghasilkan prediksi yang relatif stabil dengan risiko *overfitting* yang lebih rendah dibandingkan pendekatan klasifikasi sederhana (Dip Das, Aayushman, Sourav Kumar, Md Amir Hussain, 2025). XGBoost dan CatBoost juga memiliki mekanisme pembelajaran bertahap yang berorientasi pada peningkatan hasil prediksi sehingga dapat diterapkan pada data medis yang memiliki ketidakseimbangan distribusi kelas. Ketiga algoritma kemudian dihubungkan dengan *pipeline preprocessing* agar tahapan transformasi fitur dapat dilakukan secara otomatis dan seragam pada setiap model. Proses pembelajaran dilakukan menggunakan data latih yang terdiri atas X_train sebagai fitur masukan dan y_train sebagai target. Untuk mengurangi dampak ketidakseimbangan kelas, konfigurasi *class weighting* diterapkan pada model CatBoost dan Random Forest. Model yang telah selesai dilatih selanjutnya digunakan pada tahap pengujian untuk memperoleh nilai performa masing-masing algoritma.

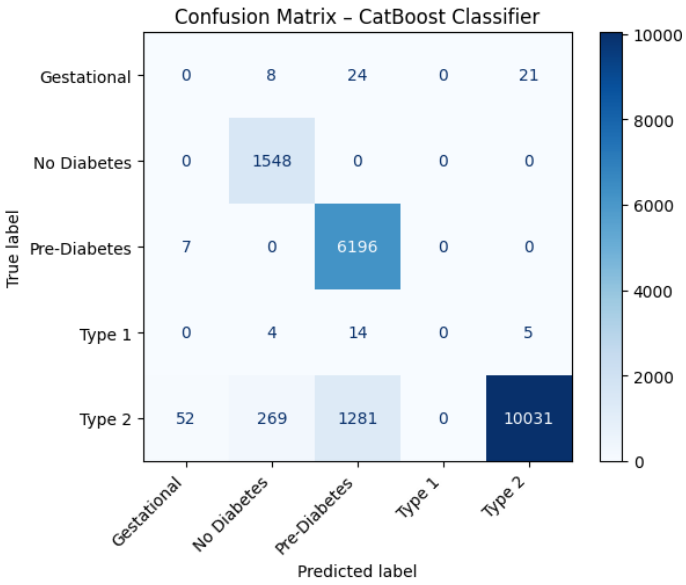
Evaluasi Model

Evaluasi dilakukan setelah proses pelatihan selesai dengan tujuan mengetahui kemampuan setiap model dalam mengklasifikasikan data. Pengukuran performa menggunakan beberapa indikator, yaitu *accuracy*, *precision*, *recall*, dan F1-*score* yang diperoleh melalui *classification report*. Selain itu, *confusion matrix* digunakan untuk memberikan gambaran mengenai jumlah prediksi yang tepat maupun kesalahan klasifikasi pada setiap kelas target. Hasil dari seluruh metrik tersebut menjadi dasar untuk membandingkan kemampuan Random Forest, XGBoost, dan CatBoost.

1. Evaluasi Model CatBoost

   CatBoost merupakan salah satu metode *gradient boosting* yang dikembangkan untuk memberikan pengolahan yang baik terhadap variabel kategorikal. Salah satu karakteristik utamanya adalah kemampuannya menangani data kategorikal dengan mekanisme internal sehingga kebutuhan terhadap proses encoding tambahan dapat dikurangi. Hal tersebut membantu model dalam mempertahankan informasi yang terdapat pada fitur kategorikal

sekaligus mengurangi potensi bias selama proses pembelajaran (Irfannandhy et al., 2024), (Linkon et al., 2024). Pada permasalahan prediksi diabetes, karakteristik tersebut menjadi salah satu pertimbangan karena dataset yang digunakan terdiri atas kombinasi fitur numerik dan kategorikal. CatBoost dapat memanfaatkan kedua jenis informasi tersebut dalam proses pembentukan model sehingga menghasilkan prediksi yang stabil. Beberapa penelitian juga melaporkan bahwa CatBoost mampu memberikan performa yang konsisten pada permasalahan klasifikasi dibandingkan sejumlah metode *ensemble* lainnya (Azizah et al., 2025).



Gambar 8. *Confusion Matrix* Catboost

Berdasarkan hasil *confusion matrix* pada model CatBoost, terdapat sebanyak 17.775 data yang berhasil diklasifikasikan dengan benar dari keseluruhan 19.460 data uji. Nilai tersebut selanjutnya digunakan untuk memperoleh tingkat akurasi model dengan menerapkan persamaan (1) sebagai berikut:

$$Accuracy = \frac{17{,}775}{19{,}460} = 0\ 9134 \qquad (1)$$

Untuk memperoleh gambaran performa model yang lebih spesifik pada masing-masing kelas, evaluasi tidak hanya dilakukan menggunakan *accuracy*. Pengukuran juga mencakup tiga metrik utama, yaitu *precision, recall, dan* F1-*Score*, yang perumusannya pada persamaan (2), (3), (4). Sebagai contoh, berikut adalah perhitungan metrik evaluasi untuk kelas Pre-Diabetes:

- *True Positive* (TP) = 6.196
- *False Positive* (FP) = 1.319
- *False Negative* (FN) = 7

$$Precision = \frac{6.196}{6.196+1.319} = \frac{6.196}{7.515} = 0\ 82 \qquad (2)$$

$$Recall = \frac{6.196}{6.196+7} = \frac{6.196}{6.203} = 0\ 99 \approx 1\ 00 \qquad (3)$$

$$F1\text{-}Score = \frac{2 \times (0{,}82 \times 1{,}00)}{0{,}82 + 1{,}00} \approx \frac{1{,}64}{1{,}82} = 0\ 90 \qquad (4)$$

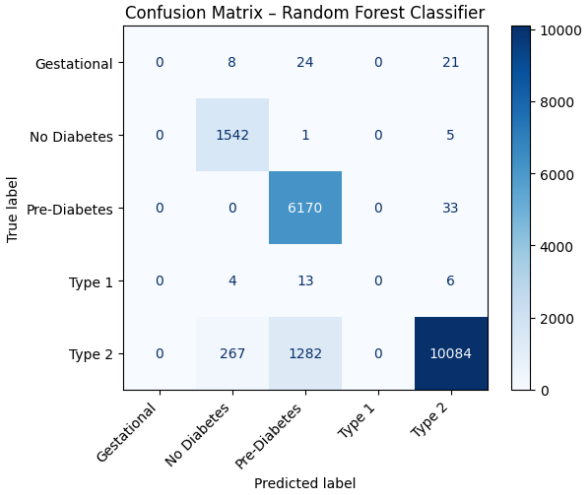
Perhitungan untuk kelas lain dilakukan dengan cara yang sama. Hasil evaluasi lengkap disajikan pada Gambar di bawah ini.

```
==========================================================
 Evaluasi Model: CatBoost Classifier
==========================================================
Akurasi: 0.9134

Classification Report:
              precision    recall  f1-score   support

 Gestational       0.00      0.00      0.00        53
 No Diabetes       0.85      1.00      0.92      1548
Pre-Diabetes       0.82      1.00      0.90      6203
      Type 1       0.00      0.00      0.00        23
      Type 2       1.00      0.86      0.92     11633

    accuracy                           0.91     19460
   macro avg       0.53      0.57      0.55     19460
weighted avg       0.93      0.91      0.91     19460
```

Gambar 9. Evaluasi Model Catboost

2. Evaluasi Model Random Forest

Random Forest termasuk metode *ensemble learning* yang menggunakan pendekatan *bagging* dengan membentuk sejumlah pohon keputusan secara terpisah. Hasil klasifikasi dari setiap pohon kemudian dikombinasikan untuk menentukan kelas akhir melalui mekanisme suara terbanyak (*majority voting*) (Setiadi et al., 2025). Penggunaan banyak pohon dalam proses pembentukan model dapat membantu menekan kemungkinan terjadinya *overfitting* sekaligus menghasilkan model yang lebih stabil dalam melakukan prediksi (Dip Das, Aayushman, Sourav Kumar, Md Amir Hussain, 2025).



Gambar 10. Confusion Matrix Random Forest

Hasil *confusion matrix* menunjukkan bahwa model Random Forest mampu mengklasifikasikan sebanyak 17.796 data dengan benar dari total 19.460 data uji. Berdasarkan jumlah prediksi tersebut, nilai *accuracy* model dapat diperoleh dengan menggunakan persamaan (1) berikut:

$$Accuracy = \frac{17{,}796}{19{,}460} = 0\ 9145 \quad (1)$$

Selanjutnya, untuk mengetahui kemampuan model dalam mengklasifikasikan setiap kelas secara lebih mendalam, digunakan tiga metrik tambahan, yaitu *precision, recall*, dan F1-*score*. Perhitungan ketiga metrik tersebut masing-masing ditunjukkan pada persamaan (2), (3), dan (4). Sebagai contoh penerapan perhitungan, evaluasi metrik dilakukan pada kelas Pre-Diabetes:

- *True Positive* (TP) = 6.170
- *False Positive* (FP) = 1.320
- *False Negative* (FN) = 33

$$Precision = \frac{6.170}{6.170+1.320} = \frac{6.170}{7.490} = 0\ 82 \quad (2)$$

$$Recall = \frac{6.170}{6.170+33} = \frac{6.170}{6.203} = 0\ 99 \quad (3)$$

$$F1\text{-}Score = \frac{2 \times (0{,}82 \times 0{,}99)}{0{,}82 + 0{,}99} \approx \frac{1{,}6236}{1{,}82} \approx 0\ 90 \quad (4)$$

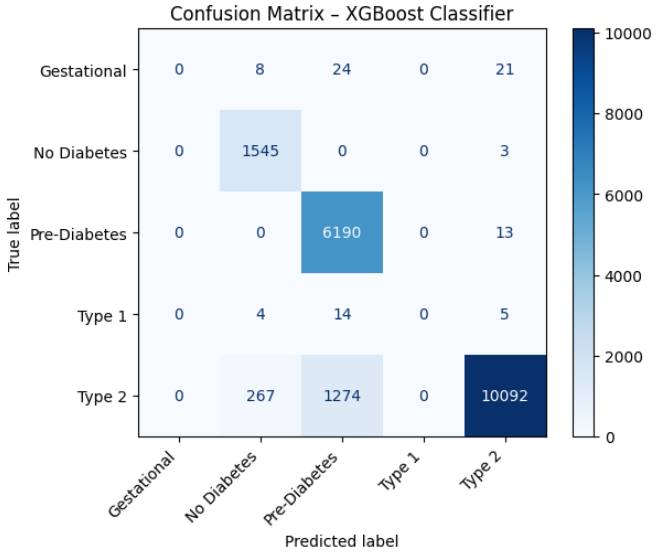
Perhitungan untuk kelas lain dilakukan dengan cara yang sama. Hasil evaluasi lengkap disajikan pada Gambar di bawah ini.

```
==================================================
 Evaluasi Model: Random Forest Classifier
==================================================
Akurasi: 0.9145

Classification Report:
              precision    recall  f1-score   support

  Gestational       0.00      0.00      0.00        53
 No Diabetes        0.85      1.00      0.92      1548
Pre-Diabetes        0.82      0.99      0.90      6203
      Type 1        0.00      0.00      0.00        23
      Type 2        0.99      0.87      0.93     11633

    accuracy                           0.91     19460
   macro avg        0.53      0.57      0.55     19460
weighted avg        0.92      0.91      0.91     19460
```

Gambar 11. Evaluasi Model Random Forest

3. Evaluasi Model XGBoost

XGBoost adalah salah satu algoritma *boosting* yang membentuk model secara berurutan, di mana setiap tahap pembelajaran diarahkan untuk memperbaiki kesalahan yang dihasilkan oleh model sebelumnya. Pendekatan tersebut memungkinkan XGBoost meningkatkan kemampuan prediksi secara bertahap. Algoritma ini juga dikenal memiliki kinerja yang baik dengan proses komputasi yang relatif efisien, sehingga sesuai untuk digunakan pada dataset berukuran besar (Tripathy et al., n.d.), (Jithendra et al., 2012).



Gambar 12. Confusion Matrix XGBoost

Berdasarkan *confusion matrix* yang diperoleh, model XGBoost mampu menghasilkan sebanyak 17.827 klasifikasi yang tepat dari keseluruhan 19.460 data uji. Nilai tersebut kemudian digunakan untuk menentukan tingkat *accuracy* model dengan menggunakan persamaan (1) berikut:

$$Accuracy = \frac{17{,}827}{19{,}460} = 0\ 9161 \quad (1)$$

Untuk memperoleh penilaian performa yang lebih terperinci pada setiap kategori, evaluasi selanjutnya dilakukan menggunakan *precision, recall*, dan F1-*score*. Ketiga metrik tersebut dihitung berdasarkan persamaan (2), (3), dan (4). Sebagai contoh, proses perhitungan nilai evaluasi ditunjukkan pada kelas Pre-Diabetes:

- *True Positive* (TP) = 6.190
- *False Positive* (FP) = 1.312
- *False Negative* (FN) = 13

$$Precision = \frac{6.190}{6.190+1.312} = \frac{6.190}{7.502} = 0\ 82 \approx 0\ 83 \quad (2)$$

$$Recall = \frac{6.190}{6.190+13} = \frac{6.190}{6.203} = 0\ 99 \approx 1\ 00 \quad (3)$$

$$F1\text{-}Score = \frac{2 \times (0{,}83 \times 1{,}00)}{0{,}83 + 1{,}00} \approx \frac{1{,}66}{1{,}83} = 0{,}90 \quad (4)$$

Perhitungan untuk kelas lain dilakukan dengan cara yang sama. Hasil evaluasi lengkap disajikan pada Gambar di bawah ini.

```
======================================================
 Evaluasi Model: XGBoost Classifier
======================================================
Akurasi: 0.9161

Classification Report:
              precision    recall  f1-score   support

 Gestational       0.00      0.00      0.00        53
 No Diabetes       0.85      1.00      0.92      1548
Pre-Diabetes       0.83      1.00      0.90      6203
      Type 1       0.00      0.00      0.00        23
      Type 2       1.00      0.87      0.93     11633

    accuracy                           0.92     19460
   macro avg       0.53      0.57      0.55     19460
weighted avg       0.93      0.92      0.92     19460
```

Gambar 13. Evaluasi Model XGBoost

```
======= RINGKASAN AKURASI MODEL =======
CatBoost     : 0.9134
RandomForest : 0.9145
XGBoost      : 0.9161
```

Gambar 14. Ringkasan Akurasi Model

Tabel 1. Tabel Komparasi Akurasi Model

Model	Kelas	Precision	Recall	F1-Score
Random Forest	No Diabetes	0.85	1.00	0.92
	Pre-Diabetes	0.82	0.99	0.90
	Type 2 Diabetes	0.99	0.87	0.93
XGBoost	No Diabetes	0.85	1.00	0.92
	Pre-Diabetes	0.83	1.00	0.90
	Type 2 Diabetes	1.00	0.87	0.93
CatBoost	No Diabetes	0.85	1.00	0.92
	Pre-Diabetes	0.82	1.00	0.90
	Type 2 Diabetes	1.00	0.86	0.92

Pembahasan

Berdasarkan Tabel 1, ketiga algoritma menunjukkan perbedaan kinerja pada masing-masing kelas. Perbandingan precision, recall, dan F1-score menunjukkan bahwa performa model tidak hanya ditentukan oleh nilai accuracy keseluruhan, tetapi juga oleh kemampuan model dalam mengenali setiap kelas risiko diabetes. XGBoost yang memperoleh accuracy tertinggi sebesar 91,61% menunjukkan performa keseluruhan yang sedikit lebih baik dibandingkan Random Forest sebesar 91,45% dan CatBoost sebesar 91,34%. Selisih accuracy yang relatif kecil menunjukkan bahwa ketiga algoritma memiliki kemampuan klasifikasi yang kompetitif pada dataset yang digunakan.

Pada tingkat kelas, perbedaan nilai precision, recall, dan F1-score menunjukkan bahwa terdapat kelas yang lebih mudah maupun lebih sulit dikenali oleh model. Perbedaan tersebut berkaitan dengan distribusi data antar kelas dan karakteristik pola kesehatan serta gaya hidup pada masing-masing kelompok. Oleh karena itu, penggunaan F1-score dan confusion matrix penting untuk melengkapi interpretasi accuracy, terutama dalam mengevaluasi kemampuan model terhadap setiap kelas secara individual.

Hasil confusion matrix juga menunjukkan bahwa sebagian besar data berhasil diklasifikasikan ke kelas yang sesuai, meskipun masih terdapat sejumlah kesalahan klasifikasi

antar kelas. Kesalahan tersebut menunjukkan adanya pola karakteristik yang saling beririsan antara beberapa kategori risiko diabetes. Berdasarkan keseluruhan metrik, XGBoost dipilih sebagai model dengan performa terbaik karena menghasilkan accuracy tertinggi serta performa klasifikasi yang kompetitif pada masing-masing kelas..

KESIMPULAN

Penelitian ini telah melakukan perbandingan terhadap tiga algoritma *ensemble learning*, yaitu Random Forest, XGBoost, dan CatBoost, dalam memprediksi risiko diabetes menggunakan dataset *Health and Lifestyle Data for Diabetes Prediction*. Evaluasi yang dilakukan berdasarkan *accuracy, precision, recall, F1-score*, serta *confusion matrix* menunjukkan bahwa seluruh algoritma mampu memberikan hasil klasifikasi yang tinggi, dengan perbedaan performa yang relatif tidak jauh. Hasil pengujian menunjukkan bahwa XGBoost memperoleh akurasi tertinggi sebesar **91,61%**, kemudian Random Forest sebesar **91,45%**, dan CatBoost sebesar **91,34%**. Berdasarkan hasil tersebut, XGBoost menjadi algoritma dengan performa paling tinggi pada dataset dan skenario eksperimen yang digunakan dalam penelitian ini. Temuan ini juga menunjukkan bahwa pendekatan *ensemble learning* dapat dimanfaatkan untuk mengenali pola pada data yang berkaitan dengan kesehatan dan gaya hidup dalam proses prediksi risiko diabetes. Dengan demikian, tujuan penelitian untuk menganalisis dan membandingkan kinerja Random Forest, XGBoost, dan CatBoost telah tercapai. Meskipun demikian, penetapan XGBoost sebagai algoritma dengan performa tertinggi masih terbatas pada kondisi eksperimen penelitian ini karena proses *cross-validation* dan *hyperparameter tuning* belum diterapkan.

Penelitian selanjutnya dapat mengembangkan hasil ini dengan melakukan optimasi *hyperparameter* dan menerapkan *cross-validation* untuk memperoleh evaluasi model yang lebih robust. Penggunaan dataset dengan karakteristik dan sumber yang lebih beragam juga dapat dipertimbangkan untuk meningkatkan kemampuan generalisasi model terhadap data baru.

REFERENSI

Afolabi, S., Ajadi, N., Jimoh, A., & Adenekan, I. (2025). Predicting diabetes using supervised machine learning algorithms on E-health records. *Informatics and Health*, *2*(1), 9–16. https://doi.org/10.1016/j.infoh.2024.12.002

Alshammari, A. F. (2024). *Implementation of Model Evaluation using Confusion Matrix in Python*. *186*(50), 42–48. https://doi.org/10.5120/ijca2024924236

Andini, S. A., Firmansyah, H., Asriyani, W., & Budiraharjo, E. (2026). *Prediksi Risiko Diabetes Mellitus Menggunakan Algoritma Neural Network*. *2*(1), 843–853. https://doi.org/10.63822/1kgj8h92

Azizah, N., Avianta, S., Putra, D. H., Satrya, B. A., & Iqbal, M. F. (2025). *Analysis of Artificial Intelligence Implementation in the Indonesian Healthcare Sector : A Literature Review*. *5*(October), 1199–1210. https://doi.org/10.57152/malcom.v5i4.2229

Dip Das, Aayushman, Sourav Kumar, Md Amir Hussain, B. R. R. (2025). Diabetes Prediction using Ensemble Learning Techniques. *Procedia Computer Science*, *258*, 3155–3164. https://doi.org/10.1016/j.procs.2025.04.573

Ibrahim, M. C., Fachruddin, & Nurhadi. (2025). *Comparison of Diabetes Prediction Data Using Machine Learning*. *5*(October), 1423–1436. https://doi.org/10.57152/malcom.v5i4.2301

Irfannandhy, R., Handoko, L. B., & Ariyanto, N. (2024). *Analisis Performa Model Random Forest dan CatBoost dengan Teknik SMOTE dalam Prediksi Risiko Diabetes*. *8*(2), 714–723. https://doi.org/10.29408/edumatic.v8i2.27990

Jithendra, V., Mohit, R. M. S., Madhusudhan, M., Jagadeesh, B., & Kusuma, S. (2012). *Diabetes Prediction using Machine Learning Techniques*. *5*(2), 190–206. https://doi.org/10.36548/jaicn.2023.2.008

Kharis Syaban, M. (2025). *Evaluasi Model Ensemble Learning pada Identifikasi Faktor Risiko Diabetes Mellitus*. *15*(September), 121–130. https://doi.org/10.34010/jati.v15i2.16238

Linkon, A. A. L. I., Noman, I. R., Islam, R., Bortty, J. O. Y. C., Bishnu, K. K., Islam, A., Hasan, R., & Abdullah, M. (2024). Evaluation of Feature Transformation and Machine Learning Models on Early Detection of Diabetes Mellitus. *IEEE Access*, *12*(October), 165425–165440. https://doi.org/10.1109/ACCESS.2024.3488743

Md Ashraful Alam ,Amir Sohel, Kh Maksudul Hasan, M. A. I. (2024). *MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE IN DIABETES PREDICTION AND MANAGEMENT : A COMPREHENSIVE REVIEW OF MODELS*. *1*(01), 107–124.

Saleh, M., Reshan, A., Amin, S., Zeb, M. A., Sulaiman, A., Shaikh, A., Member, S., & Elmagzoub,

M. A. (2024). An Innovative Ensemble Deep Learning Clinical Decision Support System for Diabetes Prediction. *IEEE Access*, *PP*(Dm), 1. https://doi.org/10.1109/ACCESS.2024.3436641

Setiadi, W. T., Jollyta, D., & Setiawan, E. B. (2025). *Prediksi Risiko Diabetes Menggunakan Model Regresi Logistik dan Random Forest*.

Shihab, A. (2026). *Health & Lifestyle Data for Diabetes Prediction*. Kaggle. https://www.kaggle.com/datasets/alamshihab075/health-and-lifestyle-data-for-diabetes-prediction/data

Tripathy, N., Moharana, B., Balabantaray, S. K., Nayak, S. K., Pati, A., & Panigrahi, A. (n.d.). A Comparative Analysis of Diabetes Prediction Using Machine Learning and Deep Learning Algorithms in Healthcare. *2024 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC)*, 1–6. https://doi.org/10.1109/ASSIC60049.2024.10508008